PSYCH-GA.2211/NEURL-GA.2201 – Fall 2017
Mathematical Tools for Neural and Cognitive Science

Homework 4

Due: 10 Nov 2017
(late homeworks penalized 10% per day)

See the course web site for submission details. Please: don't wait until the day before the due date... start *now*!

1. **Bayes' rule and eye color.** A male and female chimpanzee have blue and brown eyes, respectively. Assume a simple genetic model in which the gene for brown eyes is always dominant (so that the trait of blue eyes can only arise from two blue-eyed genes, but the trait of brown eyes can arise from two brown-eyed genes, or one of each). You can also assume that the apriori probability that the female has either of these gene configurations is 50%.

   (a) Suppose you observe that they have a single child with brown eyes. What is the probability that the female chimp has a blue-eyed gene?

   (b) Suppose you observe that they have a second child with brown eyes. Now what is the probability?

   (c) Generalizing, suppose they have N children with brown eyes... express the probability, as a function of N.

2. **Sums of random variables.** Consider a discrete distribution specified by a vector `p` of length `n` whose elements provide the frequency of occurrence of each of the integers $1 \ldots n$. (The values in `p` should be non-negative and sum to 1).

   (a) Write a function `samples = randp(p, num)` that generates `num` samples from the PDF specified by `p`. Test your function by choosing some arbitrary `p` of length 10, drawing 1,000 samples, plotting a histogram of how many times each value is sampled, and comparing this to the frequencies predicted by `p`.

   (b) Next, write a function `psum(p, q)` that, given two such discrete PDFs, returns a vector encoding the PDF for the sum of a sample drawn from `p` and a sample drawn from q. (Hint: the output vector should have length $m + n - 1$ when m and n are the lengths of the two input PDFs.)

   Test your function by letting `p = [1:6] / sum([1:6])` (a weighted die), and using repeated calls to `psum` to compute the PDF predicted for a sum of five rolls of the die. Use your function `randp` to generate 5 sets of 1000 samples from `p`; plot a histogram of how many times each sum is sampled; and compare this the frequencies predicted by your program.

3. **The Central Limit theorem.** The Central Limit theorem states that the distribution of the average of a set of samples (drawn independently, from any distribution with finite mean and variance) gets closer and closer to a Normal (Gaussian) distribution as the size of the sample increases. Specifically, if the mean and variance of the original distribtion are μ and σ, the distribution of the average converges to $\mathcal{N}(\mu, \sigma/\sqrt{n})$ as n increases.

(a) Generate 1,000 samples of two values each from a uniform distribution (use `rand`). Compute the average of each sample (pair of values), and plot a histogram of these. What shape is it, approximately? Why? If you're unsure, try it again with more samples (e.g., 100,000).

(b) Now try this again with samples containing 3 values. How has the histogram changed? Try sample sizes of 4 and 5 as well. When do you judge that the histogram starts looking Normal?

(c) Let's test the Normality of the distribution a bit more carefully, using a "Q-Q" (quantile-quantile) plot, which plots the quantiles of one distribution against another. If the two distributions match, the values should lie on a unit-slope line. For this problem, you can use the function `normplot`, which plots the quantiles of a sample of data against those of a Normal distribution of the same mean and variance. First, try this on a sample of 1,000 values from a normal distribution (use `randn`). The points should fall (close to) a straight line, indicating that the sample is close to normal, as expected. Try this a few times to see how the plot varies (you might want to put them on the same graph, using matlab's `hold on` command).

(d) Now call `normplot` on a sample of 1,000 values from a uniform distribution. Explain qualitatively why it has the shape it does (hint: think about the quantiles of the uniform and Normal distributions). Do this for averages of uniform samples of different size (2, 3, 4, ...). Keep increasing sample size until you cant tell the resulting QQ plot from the QQ plots for samples from a normal distribution. How big does the sample have to be?

4. **Multi-dimensional Gaussians.**

(a) Write a function `samples = ndRandn(mean, cov, num)` that generates a set of samples drawn from a multidimensional Gaussian distribution with the specified `mean` (an N-vector) and `cov`ariance (an NxN matrix). The parameter `num` should be optional (defaulting to 1) and should specify the number of samples to return. The returned value should be a matrix with `num` rows each containing a sample of N elements. (Hint: use the MATLAB function `randn` to generate samples from an N-dimensional Gaussian with zero mean and identity covariance matrix, and then modify these appropriately. Recall that the covariance of $Y = MX$ is $E(YY^T) = MC_XM^T$ where C_X is the covariance of X).

(b) Test your function by plotting 1000 samples of a 2-dimensional Gaussian (choose an arbitrary nonzero mean and nonzero covariance). Measure the sample mean and covariance of your data points, comparing to the values that you requested when calling the function. Plot an ellipse on top of the scatterplot that traces out points that are two standard deviations away from the mean, according to the covariance matrix. Does this ellipse capture the shape of the data?

(c) Now consider the generalized marginal distribution of your 2-D Gaussian in which samples are projected onto a unit vector $\vec{u}$ to obtain a 1-D distribution. Write a mathematical expression for the mean and variance of this marginal distribution as a function of $\vec{u}$ and check it for a set of 48 unit vectors spaced evenly around the unit circle. For each of these, compare the mean and variance predicted from your mathematical expression to the sample mean and variance estimated by projecting your 1,000 samples onto $\vec{u}$. Plot the mathematical mean and the sample mean (on the same plot), and also plot the mathematical variance and the sample variance.